DOCUMENT RESUME

ED 100 718 SE 018 749

AUTHOR Zalewski, Donald L.
TITLE An Exploratory Study to Compare Two Performance Measures: An Interview-Coding Scheme of Mathematical Problem Solving and a Written Test. Part 1. Technical Report No. 306.
INSTITUTION Wisconsin Univ., Madison. Research and Development Center for Cognitive Learning.
SPONS AGENCY National Inst. of Education (DHEW), Washington, D.C.
REPORT NO WRDCCL-TR-306
PUB DATE Aug 74
CONTRACT NE-C-00-3-0065
NOTE 99p.; Report from the Project on Conditions of School Learning and Instructional Strategies. For Part 2, see SE 018 750

EDRS PRICE MF-$0.75 HC-$4.20 PLUS POSTAGE
DESCRIPTORS Cognitive Measurement; *Cognitive Objectives; Cognitive Tests; Doctoral Theses; Grade 7; *Mathematics Education; *Problem Solving; *Research; Secondary School Mathematics; Tests; *Test Validity

ABSTRACT

Observing that available standardized tests purporting to measure problem solving in mathematics do not validly reflect currently accepted definitions of problem solving, while more valid individualized measures are costly in terms of time, this researcher undertook (1) to construct a paper-and-pencil instrument which might better reflect problem-solving ability and (2) to validate this test by correlation with scores of subjects in taped and coded problem-solving interviews. Seventh-grade students with average and above-average mathematics achievement records were given two written tests in a group situation; later each was interviewed and given six problems to solve while talking aloud in a videotaping or audiotaping situation. Scores were assigned to performance in the individual session according to a previously developed coding system. Test statistics were computed; differences in the rank ordering of subjects by the two measures and correlations among scores were analyzed; differences in performance and coding ease between videotaped and audiotaped subjects were investigated. Results and instruments used are reported in Part 2 of this report. (SD)

U.S. DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

SCOPE OF INTEREST NOTICE
The ERIC Facility has assigned this document for processing to: SE TM

In our judgement, this document is also of interest to the clearinghouses noted to the right. Indexing should reflect their special points of view.

SE 018 199

ED100718

Technical Report No. 306 (Part 1 of 2 Parts)

AN EXPLORATORY STUDY TO COMPARE TWO PERFORMANCE MEASURES: AN INTERVIEW-CODING SCHEME OF MATHEMATICAL PROBLEM SOLVING AND A WRITTEN TEST

Report from the Project on Conditions of School Learning and Instructional Strategies

By Donald L. Zalewski

Thomas A. Romberg and John G. Harvey
Principal Investigators

Wisconsin Research and Development
Center for Cognitive Learning
The University of Wisconsin
Madison, Wisconsin

August 1974

Published by the Wisconsin Research and Development Center for Cognitive Learning, supported in part as a research and development center by funds from the National Institute of Education, Department of Health, Education, and Welfare. The opinions expressed herein do not necessarily reflect the position or policy of the National Institute of Education and no official endorsement by that agency should be inferred

Center Contract No. NE-C-00-3-0065

STATEMENT OF FOCUS

Individually Guided Education (IGE) is a new comprehensive system of elementary education. The following components of the IGE system are in varying stages of development and implementation: a new organization for instruction and related administrative arrangements; a model of instructional programing for the individual student; and curriculum components in prereading, reading, mathematics, motivation, and environmental education. The development of other curriculum components, of a system for managing instruction by computer, and of instructional strategies is needed to complete the system. Continuing programmatic research is required to provide a sound knowledge base for the components under development and for improved second generation components. Finally, systematic implementation is essential so that the products will function properly in the IGE schools.

The Center plans and carries out the research, development, and implementation components of its IGE program in this sequence: (1) identify the needs and delimit the component problem area; (2) assess the possible constraints--financial resources and availability of staff; (3) formulate general plans and specific procedures for solving the problems; (4) secure and allocate human and material resources to carry out the plans; (5) provide for effective communication among personnel and efficient management of activities and resources; and (6) evaluate the effectiveness of each activity and its contribution to the total program and correct any difficulties through feedback mechanisms and appropriate management techniques.

A self-renewing system of elementary education is projected in each participating elementary school, i.e., one which is less dependent on external sources for direction and is more responsive to the needs of the children attending each particular school. In the IGE schools, Center-developed and other curriculum products compatible with the Center's instructional programing model will lead to higher morale and job satisfaction among educational personnel. Each developmental product makes its unique contribution to IGE as it is implemented in the schools. The various research components add to the knowledge of Center practitioners, developers, and theorists.

ACKNOWLEDGMENTS

Since this dissertation is the culmination of many years of formal learning, I am deeply indebted to the people who provided the opportunities for my education and to the people who helped me through the investigation. Although there is no way to repay them, I take this opportunity to say "Thank You!"

My first thanks go to the grade school, high school, and college teachers who made learning enjoyable and who encouraged me to continue my education. Thanks also to relatives and friends whose confidence in me often exceeded my own.

I owe special thanks to Professors Thomas A. Romberg, John G. Harvey, and J. Fred Weaver for the uncounted formal and informal occasions when I benefited from their expertise, opinions and advice. My major advisor, Dr. Romberg, was instrumental in formulating this study and was a constant source of encouragement and assistance. Professor Harvey gave generously of his time throughout the study and his editorial comments helped improve my writing style and the readibility of the thesis. Professor Weaver raised constructive questions during the study and served as the third member of the thesis committee.

I also wish to thank Professors Elizabeth H. Fennema and Larry J. Hubert for reading the thesis and contributing their opinions and suggestions.

The Wisconsin Research and Development Center for Cognitive Learning provided the equipment, facilities, and staff to conduct the study and complete this report. In particular, Mr. John McFee provided the technical expertise and manpower for the video taping, Misses Linda Junker and Pat Busk assisted in the data analyses, and Mrs. Ethel Koshalek typed the final draft. I am grateful for their assistance.

Professor John F. Lucas, presently teaching at the University of Wisconsin in Oshkosh, helped me to administer his coding scheme. Mr. Norman Loomer and Sister Ruth Meyer worked diligently as additional coders during the reliability tests. Their willing help made the task easier and I appreciate their efforts.

I am particularly grateful to the students who participated in the study and the teachers who allowed me to disrupt their schedules. Their cooperation made scheduling easy and provided the interesting data and observations recorded herein.

Finally, I thank the people who were most responsible for this dissertation: my parents, wife, and children. As my first teachers, my parents, instilled in me a patience and a faith which proved invaluable. My wife, Rita, encouraged me to continue the studies which resulted in this thesis and shared the moments of inspiration, perspiration, and desperation with me. Our children, Tami and Stephen, patiently understood their father's frequent absence and inattention during the research and the preparation of the manuscript. For their contributions and their confidence in me, I gratefully dedicate this thesis to my parents, wife, and children.

TABLE OF CONTENTS

		Page
ACKNOWLEDGMENTS		iv
LIST OF TABLES		xi
LIST OF FIGURES		xiii
ABSTRACT		xv
CHAPTER		
I	INTRODUCTION	1
	Definitions	2
	Outline of Thesis	5
II	THE PROBLEM	7
	Introduction	7
	General Problem	7
	The Need for Problem Solving Measurement Procedures	7
	Limitations of Existing Procedures	11
	Specific Problem	14
	Summary of Chapter II	17
III	REVIEW OF RESEARCH AND LITERATURE	19
	Introduction	19
	General Problem Solving	19
	Past Problem Solving Research	20
	The Thinking Aloud Procedure	22
	Mathematical Problem Solving	24
	Problem Variables	24
	Coding Schemes	26
	Test Criteria and Construction	28
	Test Reliability	29
	Test Validity	29
	Validity of Available Procedures	32
	Summary of Chapter III	39

CHAPTER		Page
IV	DESIGN OF THE STUDY	41
	Introduction	41
	Part I: The Complex Problem Solving Assessment Procedure	41
	The Mathematical Problems	42
	The Subjects	43
	The Interviews	44
	The Coding System	46
	The Ranking	47
	Summary of Part I	48
	Part II: The Written Test	49
	The WT Items	49
	Administration of the WT	51
	The Ranking	52
	Summary of Part II	55
	Part III: The Comparison of Ranks	55
	Correlations	56
	Processes and Patterns	57
	Testing Procedures	57
	Interpreting Results	59
	Summary of Chapter IV	61
V	EXECUTION OF THE PLANS	63
	Introduction	63
	Pilot Study	63
	Pilot Study Sample	63
	Audio Taping Versus Video Taping	64
	Changes in the Written Test	71
	Changes in the Interview Procedures	71
	Changes in the Coding System and Checklist	73
	Main Study	75
	Item Pools and Samples	75
	Population	76
	Written Test Administration	76
	The Interview Sample	78
	The Interview Arrangements	79
	The Interview Proceedings	81
	Summary of Chapter V	84

CHAPTER		Page
VI	DATA AND ANALYSES	85
	Introduction	85
	The Written Test (WT)	85
	Subject Response Data	85
	WT Length and Reliability	89
	Written Test Rankings	91
	The Interview Test (IT)	93
	The Thinking Aloud Procedure	93
	The Coding Systems	97
	The IT Ranking Schemes	105
	Audio Versus Video Taping	109
	Statistical Analyses of Rankings	115
	Relationships of the Written and Interview Tests	115
	Exploratory Procedures	117
	Summary of Chapter VI	124
VII	CONCLUSION	127
	Introduction	127
	Summary	127
	Limitations	128
	Conclusions	130
	Implications for Mathematical Problem Solving Assessment	133
	Recommendations for Future Research	136
	Comments	139
REFERENCES		141
APPENDIX A:	KILPATRICK'S CODING FORM FOR PROBLEM-SOLVING PROTOCOLS	147
APPENDIX B:	LUCAS' PROCESS-SEQUENCE CODES	144
APPENDIX C:	INTERVIEW TEST ITEM POOL WITH ANSWERS	153
APPENDIX D:	WRITTEN TEST ITEMS WITH ANSWERS	161
APPENDIX E:	INSTRUCTIONS FOR THE INTERVIEW TEST	179
APPENDIX F:	SUMMARY OF LUCAS' SCORING SYSTEM	181
APPENDIX G:	PROCESS-SEQUENCE CODES	183

Page

APPENDIX H: PILOT STUDY WT RESULTS 187

APPENDIX I: SOLUTION AND CODING TIMES OF SUBJECTS' PROTOCOLS 189

APPENDIX J: AGREEMENT ON CODING AND SCORING VARIABLES 191

APPENDIX K: SUBJECT SCORES ON THE INTERVIEW TEST . 195

APPENDIX L: SIMILARITY MEASURES. 199

APPENDIX M: MULTIDIMENSIONAL SCALING RESULTS FOR 2, 3, AND 4 DIMENSIONS 203

APPENDIX N: GAMMA VALUES FOR CLUSTERING 205

LIST OF TABLES

Table		Page
6.1	Mean, Standard Deviation, and Range for the WT: Group A, Group B, and Combined	86
6.2	Descriptive Statistics for the WT2: By School, Combined, and by Groups A and B	88
6.3	Hoyt's Reliability Coefficient for the WT and WT2	90
6.4	Rankings of Group A Based on the Results of the WT and the WT2.	92
6.5	Indicators of Thinking Aloud Difficulties	94
6.6	Thinking Aloud Rating of Subjects	96
6.7	Agreement Measure Averages Over Coders 1, 2, and 3.	102
6.8	Interview Test Scores and Rankings A, B, and C. . .	106
6.9	Analysis of Variance for Total Interview Test Scores	111
6.10	Analysis of Variance for Subjects' Total Solution Times on the Interview Test	112
6.11	Analysis of Variance for Coding Times	113
6.12	Comparison of Coding Time Ratios	114
6.13	Correlation and Ranking Statistics for the Interview Test and the Written Tests	116
6.14	One Dimensional Scaling Coordinates and a Resulting Ranking	120

LIST OF FIGURES

Figure		Page
5.1	One Way Fixed Effects ANOVA of Process Scores . . .	68
5.2	One Way Fixed Effects ANOVA of Solution Times . . .	68
5.3	Two Ways to Test for Coding Time Differences . . .	70
6.1	Distribution of the Numbers of Correct Responses on the Written Test (WT)	87
6.2	Similarity Measure Formula	118
6.3	Clustering Algorithm Graph	122

ABSTRACT

BEST COPY AVAILABLE

The investigation reported in this thesis is on assessment in mathematics education. Specifically, this study explored the feasibility of using a written test to predict seventh graders' mathematical problem solving achievement as assessed by an interview-coding procedure.

A search revealed that most available mathematical problem solving assessment procedures are commercial tests. The tests do not offer any definitions and their items are usually simple applications or algorithmic situations which do not satisfy the criteria established in this thesis for a mathematical problem.

The method for validly assessing subjects' mathematical problem solving achievement used in this study was a thinking aloud procedure. Interviews yielded audio and video taped protocols, and a coding system permitted classification, analysis, and scoring of the subjects' performances. Because of the complexity of the interview and coding scheme, a written instrument which hopefully had high concurrent validity was developed so that it could be used as a valid alternative to the interview and coding procedure.

Thirty-one seventh graders were asked to think aloud as they tried to solve six mathematical problems in individually taped interviews. The subjects' protocols were coded and scored to provide what

was assumed to be a valid assessment of their mathematical problem solving achievement. The 31 subjects also took two 20 item written tests which were scored by the number of correct responses. Three rankings were developed from the interview test and one ranking was developed from each written test.

The correlation coefficients between the written and interview test scores did not reach the .71 level established for feasibility. One coefficient reached .68 and the tests shared high rank order agreement. These results suggested that a more reliable test might attain the .71 correlation. Clustering and multidimensional scaling verified the structure imposed by the total score ranks.

Other findings indicated that present coding schemes can be applied reliably to describe subjects' problem solving behaviors and that the scoring system permits logical ranking of the subjects. However, serious questions were raised about the validity of the thinking aloud procedure. Video taping the interviews was advantageous because it captured silent indicators of problem solving behaviors and took less time to code.

Chapter I

INTRODUCTION

BEST COPY AVAILABLE

Major mathematics curriculum revision proposals, projects developing mathematics curricula, and individual school programs prominently include emphasis upon problem solving skills in their goals. Consequently, the means of enabling students to develop good problem solving techniques and the assessment of problem solving achievement have been the objectives of research investigations. Progress in the development of instruments for analyzing problem solving has been slow, but the importance placed upon problem solving and the presently inadequate understanding of this complex behavior demand continued probing. This investigation will continue the efforts of earlier studies which attempted to analyze mathematical problem solving performances of students.

Many psychological investigations have studied problem solving, but only a small minority of them have concentrated on mathematical problem solving. A subset of these problem solving studies in mathematics have directed their attention to the school environment, and they are further divided into those which attempted to promote the development of problem solving behaviors and those which attempted to develop instruments for analyzing the behaviors of students in problem

solving situations. The latter purpose formed the focal point of this investigation. The available procedures for observing and classifying students' behaviors in mathematical problem solving situations were examined. "Thinking aloud" was accepted as a valid procedure for having subjects externalize their problem solving processes and this method was used as subjects solved mathematical problems in an interview test (IT) situation. An available coding scheme was adapted to the sophistication of the recorded protocols. After adjustments and refinements, the revised procedures were used with a sample of students to classify and rank their performance. Finally, the feasibility of devising a simpler instrument, in particular a written test (WT), which produced a similar ranking was explored.

The contributions and limitations of previous research in problem solving will be discussed in Chapter III. However, in order to reduce the scope of the survey, the definitions to be used in this study are introduced.

Definitions

The content of any problem solving study depends on its interpretation of the term "problem." In general, a problem is a situation which presents an objective that an individual is motivated to achieve but for which he has no immediate procedures to arrive at the objective. This definition includes the entire gamut of problems ranging

from complex personal decisions involving combinations of emotions or facts to well defined tasks such as seeking a pattern in a series of numbers. The difficulty of identifying and studying problem solving behaviors involved in such a variety of situations has resulted in sporadic development of this research area. Lucas (1972) summarizes the reasons for the slow progress:

> Consequently, the pertinent literature of psychology and education is replete with semantic ambiguities, isolated task situations, inferences from observables to unobservables, lack of consolidation of research effort, and a host of other characteristics which serve to retain in a somewhat primitive state a field which has been considerably researched. This is not to deplore the existing state of research on problem solving, but rather to point out that the complex nature of the subject practically demands that progress occur most frequently by small steps and only occasionally by giant leaps. (Lucas, 1972, p. 6-7)

In order to avoid some of the complexities of research in problem solving and to gain more control over the variables of interest, investigators have placed numerous limitations on their studies. They have used restricted definitions of "problem," have looked at the simple forms of problem solving, and have introduced manipulative tasks to produce more observable responses. These restraints undoubtedly distort uninhibited problem solving behavior, but they are necessary to continue the investigative progress being made in this area.

The mathematics educator is confronted with additional difficulties

when attempting to investigate problem solving. Lucas (1972) explains, "Unfortunately for the mathematics educator, the nature of tasks used in most psychological investigations is not typical of the kind encountered in the study of mathematics (1972, p. 7)." Complex, challenging mathematical problems rarely appear in research literature. Kilpatrick (1967) gives three reasons for this neglect:

> Reasons for such neglect are not hard to find. Performance on any problem depends critically on the nature and extent of one's appropriate past experience, and in the case of complex mathematical problems, this factor is difficult to control. Most challenging problems, moreover, take so long to solve that the investigator risks tiring or boring his subjects before he can get stable measures of behavior. Finally, and perhaps most importantly, subjects exhibit little observable behavior while solving a mathematical problem in comparison with a problem requiring the manipulation of apparatus. Given these difficulties and the premium put on experimentation in psychological research, it is not too surprising that investigators have chosen to use simple tasks in studying problem solving. (Kilpatrick, 1967, p. 2)

The complications and limitations of previous problem solving research have influenced the definition which was formulated for this study. A mathematical problem is a written or oral statement which meets three conditions: 1) the statement presents information and an objective or a question whose answer is based on the information; 2) the objective or answer can be found by translation of the information into

BEST COPY AVAILABLE

mathematical terms and/or by application of rules from mathematical areas as arithmetic, algebra, logic, reasoning, geometry, number theory, or topology; and 3) the individual attempting to answer the question or attain the objective does not possess a memorized answer or an immediate procedure. If an individual solved a given problem or a similar one previously and simply recalls the answer or appropriate procedure, the situation would not be considered a problem for him. The particular restrictions on the mathematical problems to be used in this investigation will be discussed in Chapter IV.

A second definition is included before an outline of the remaining chapters is presented. Mathematical problem solving is the process of producing and using a procedure to achieve the solution to a mathematical problem. The process involved in solving the problem may require a search of possible strategies, the use of various rules and techniques, and prior knowledge of the areas mentioned in the definition of a mathematical problem.

The two definitions given above are the only ones which are unique to this study. They are the basis for the discussions and plans of the remaining chapters of the thesis.

Outline of Thesis

Chapter II stresses the significance of the general problem and development of the specific problem as the rationale and purpose of this study are discussed. The contributions and limitations of

relevant research are examined in Chapter III.

Chapter IV details the proposed procedures, items, and design of the study. The results and implications of a pilot study are included in Chapter V prior to the accounts of the main study. Chapter VI provides the data and analyses of the study and Chapter VII, the final chapter, interprets the data within the limitations of the study. A summary and the implications for further research conclude the discussion of the study.

Chapter II

THE PROBLEM

Introduction

While engaged in developing test items for a statewide mathematics assessment program, the investigator realized that very few methods to record and assess the mathematical problem solving achievement of students exist. This exploratory study was designed to investigate the existing procedures for measuring mathematical problem solving skills and to explore the feasibility of using a short written test to assess mathematical problem solving achievement. In this chapter, the status and significance of the general problem is reported and the specific problem is identified. The particular questions which must be answered are followed by a description of planned research strategies.

General Problem

The Need for Problem Solving Measurement Procedures

Curriculum developers and implementers who are interested in assessing problem solving achievement in mathematics need valid and reliable measuring procedures. A search by the investigator produced only a few procedures which claim to measure such achievement, and an examination (detailed in the next section) of these procedures raised doubts of their validity. For example, the Iowa

Test of Basic Skills (Lindquist and Hieronymus, 1964), Form 2, is identified as a problem solving instrument, but the items would not satisfy the definition of mathematical problem used in this study because direct algorithmic processes are suggested by words like "total" and "difference." The examination of other available instruments and procedures raised similar questions of validity and stressed the need to develop problem solving assessment methods.

The need to develop assessment procedures is magnified by the importance attached to problems and problem solving skills by groups and individuals concerned with the mathematics curriculum. In its recommendation Program for College Preparatory Mathematics, the College Entrance Examination Board (CEEB) stated, "Throughout the teaching of the entire sequence, the Commission urges that the mathematics taught be applied to problems, real or puzzle type (CEEB, 1959, p. 35)." Fehr stressed problem solving as he discussed the goals of school mathematics instruction:

> Our (the mathematics education community) instruction serves to develop the capacity of the human mind for the observation, selection, generalization, abstraction, and construction of models for use in solving problems in other disciplines. Unless the study of mathematics can operate to clarify and to solve human problems, it has indeed only narrow value. (Fehr, 1974, p. 27)

Polya (1962) expressed a similar but stronger opinion:

> What is know-how in mathematics? The

> ability to solve problems - not merely routine problems but problems requiring some degree of independence, judgment, originality, creativity. Therefore, the first and foremost duty of the high school in teaching mathematics is to emphasize methodical work in problem solving. (Polya, 1962, viii)

If educators attempt to develop the mathematical thinking suggested by Fehr and Polya or to follow the recommendations of the CEEB, then problem solving assessment procedures will be needed to judge the success of their efforts.

Another important need for problem solving assessment methods has resulted from recent trends in state accountability testing. States are assessing the capabilities of their students with commercial tests or their own evaluative instruments. In either case, the tests cover a multitude of subject areas, skills, attitudes, and characteristics. Mathematical problem solving is often mentioned, but rarely included. Wisconsin included "problem solving" situations in its 1973 state mathematics assessment in order to get a measure of children's ability to apply mathematical ideas. (Thompson, 1974). Minnesota's state assessment committee in mathematics listed problem solving as a desirable outcome of instruction and identified skills and problems to be included in the tests (Dye, 1972). Despite the importance attached to this area, few tests on problem solving in mathematics are available. The creation of a new instrument is a formidable task and a specific assessment tool may not satisfy the needs

of all states, but an example or model would be helpful as a guide to future test designers.

The most direct and practical need for problem solving measurement procedures arises in the classroom. Teachers who promote mathematical problem solving need methods to measure the achievement and assess the progress of their students. Instructors have some procedures for evaluating the students' problem solving skills, but are prone to use only final answers as the criteria for evaluating student achievement. Partial credit is given for evidence of proper procedures in written work, but observation of a student's answers or writing reveals little of the processes, hesitations, and difficulties which may have occurred. Judgment of the student's efforts are based on inferences and speculation of the behaviors which may have prevailed. An improved method of recording and assessing students' achievement should reduce the possibility of error by furnishing reliable evidence of the behaviors involved.

The need for procedures which identify students' problem solving behaviors and assess the status of their achievement is a result of the emphasis given to problem solving in the mathematics curriculum. This need led the investigator to seek and examine the available methods. It was found that the mathematical problem solving assessment procedures available to schools were commercial written instruments. An inspection of these tests revealed inadequacies and resulted in an examination of research procedures. The examination of exist-

ing products and procedures is reported to substantiate the general problem and establish the specific purpose of this study.

Limitations of Existing Procedures

Existing methods for measuring mathematical problem solving achievement were inspected and it became apparent that a commercial test was the type of procedure which was available to schools. The available tests were examined for the existence of a mathematical problem solving subtest and the items and design of such subtests were examined. Several inadequacies in both the items and scoring procedures were detected in the instruments.

"Mathematical problem solving" is one of the tests in the Metropolitan Achievement Test (Durost, et al., 1962) batteries, but the items are simple verbal situations often necessitating only one obvious operation through questions like "How much more ...?", "How many times as many ...?", or "What is the area of ...?". In items requiring two operations and more complex behaviors, the students need only select the appropriate sentence from four choices (the fourth being "more information needed"). The simple situations which can be solved with the direct application of a single suggested algorithm would not demand the search for an appropriate procedure that the definition of mathematical problem used in this study requires. The complex situations do not qualify as mathematical problems since a choice of alternatives was given and the final solution to the item was not sought.

The ITBS mentioned earlier in this chapter, contained "problem solving" items similar to those of the MAT. The existence of clues such as "total" and a single step algorithmic solution to an item made the items unacceptable as mathematical problems.

The Instructional Objectives Exchange (IOX. 1970) identifies a major category, "Application--Problem Solving." The questions give attention to both process and solution, but the sample objectives emphasize the solution of the items and a student is rated by the number of correct answers he chooses.

The IOX and all the commercial tests which the investigator examined present a choice of answers (usually four or five) for each item. Though this practice permits rapid scoring, it does not create a genuine problem solving situation. The search for procedures and possible solutions is an essential part of this study's definition of mathematical problem solving. Thus, any mathematics test which provides alternatives does not qualify as a valid problem solving instrument.

After commercial tests were examined and judged inadequate, research practices were investigated for mathematical problem solving assessment procedures. The most promising development being used by researchers was a thinking aloud procedure. It asks subjects to verbalize their thoughts as they solve problems and obtains a record of observable behavior. The accepted usage of a thinking aloud

procedure in research underlies the assumption that the resulting subject protocols provide valid insights into problem solving behaviors.

A thinking aloud procedure has been made particularly amenable to mathematical problem solving research through the efforts of Kilpatrick (1967) and Lucas (1972). Kilpatrick produced a valuable guide for coding protocols which are audio taped during problem solving sessions in mathematics. Lucas modified Kilpatrick's classification system during a study involving heuristic problem solving strategies in calculus. The combination of the thinking aloud procedure and Lucas' refined classification scheme has developed into a valuable instrument for assessing mathematical problem solving achievement.

The use of protocols resulting from thinking aloud sessions, followed by analysis and evaluation based on a coding scheme, was assumed to be a valid method of classifying students' mathematical problem solving performances. However, the method is not readily used because of its physical limitations. The immediate disadvantages of the thinking aloud procedure are obvious: only one subject can be tested at a time; considerable time and expense are involved in observing, coding, and evaluating each performance; and specially trained interviewers and coders are needed. These factors would make a large scale assessment financially impractical, if not impossible for schools or school districts, and also for states with restricted

budgets. For the individual teacher, the lack of interview and coding skills could be a deterrent, and finding additional time for interviews in already crowded schedules makes the thinking aloud procedure prohibitive for classroom use. The physical limitations of this valid research procedure combined with the inadequacies of available written instruments underlie the general problem of the study.

The development of the general problem can be summarized as follows: A) Problem solving is an important element of the mathematics curriculum; B) Methods for measuring and assessing mathematical problem solving achievement are needed; C) The thinking aloud procedure and coding scheme is a valid procedure for recording and assessing a student's problem solving skills; and D) The thinking aloud procedure is impractical for ordinary use in schools. The physical limitations of the thinking aloud procedure and the absence of valid alternatives for measuring problem solving fostered the specific purpose of this study.

Specific Problem

Research has produced an acceptable method for validly examining mathematical problem solving behaviors. The thinking aloud procedure removes the necessity of making speculative inferences about a student's procedures and captures the verbalized processes a student has used. Coding schemes have made it possible to classify and assess

the behaviors of subjects, but the time and training required by the interview and coding system make it impractical for use in schools. This study was devised to seek a practical alternative for measuring mathematical problem solving achievement.

A practical alternative for measuring problem solving performances is a paper and pencil test, as it requires only simple materials and does not necessitate specially trained personnel to administer it. This study investigated the feasibility of producing a written instrument that reflects the mathematical problem solving ability of seventh grade students. Seventh grade was chosen in order to reduce the scope of this study and to utilize Kilpatrick's (1967) suggestions for working with students near this level.

The development of a written instrument for assessing the mathematical problem solving achievement of seventh graders was dependent upon the results of thinking aloud interviews and coding systems, thus it necessitated answers to other questions. Specifically, three questions were regarded:

1) How well does the thinking aloud procedure and related coding scheme capture and classify the mathematical problem solving behaviors of seventh graders?
2) Is it possible to assess, separate, and rank seventh graders according to their coded problem solving protocols?

3) Is it feasible to construct a written evaluative instrument whose results correlate well with the ranking derived from the coded protocols?

The details of how answers to these questions were sought are discussed in Chapter IV while the general plans for this study are given below.

In order to avoid amibguous or inappropriate mathematical problem solving items as found in the commercial tests, a definition of mathematical problems was established and a large pool of representative items created. Seventh graders were tested on a sample of the representative mathematical problems during individual thinking aloud interviews and their protocols were coded. It was assumed that a valid record of problem solving behaviors was now obtained. Necessary adjustments in the interview or coding procedure were used to answer question number 1.

A paper and pencil instrument was created to provide a ranking of the same students who participated in the interview. Finally, an assessment and ranking scheme for the coded protocols of the subjects was developed. A search for a high correlation between the two ranks included adjustments in the exploratory procedures and provided information for questions 2 and 3.

The paper and pencil instrument which resulted from such procedures would not be classified as a problem solving test nor would the items

be typical mathematical problems. However, if its results correlate well with the complex mathematical problem solving assessment system, then the paper and pencil test should provide a reliable ranking of students' problem solving achievement and perhaps some insights into their behaviors.

Summary of Chapter II

Products and procedures which claim to assess mathematical problem solving achievement are available, but their validity is suspect. The thinking aloud procedure produces a valid record of a subject's behaviors, but is impractical for most school applications. This study explored the feasibility of producing a written test which has concurrent validity with the thinking aloud procedure. A review of previous research in mathematical problem solving is reported in Chapter III.

Chapter III

REVIEW OF RESEARCH AND LITERATURE

Introduction

Research has made sporadic progress in analyzing problem solving behaviors and achievement. However, earlier studies have produced valuable results and guidelines which have been incorporated into this study. The benefits were derived from three major areas of research: general problem solving, mathematical problem solving and test construction. The relevant literature and the resulting directions of these three areas are reported in this chapter.

In addition to the studies on problem solving and test construction, the existing instruments for measuring problem solving achievement are examined and reported. The limitations of the tests are combined with the research directives to produce the guidelines for this study. General problem solving research is reported first, followed by the studies from mathematical problem solving and test construction. The guidelines and results adopted from the reports are summarized to conclude this chapter.

General Problem Solving

Psychologists have long been interested in problem solving. At the turn of the century, Ruger (1910) discussed the solution of mechanical puzzles and the acquisition of skills in their manipulation.

Other investigators have since reported on the numerous aspects of problem solving such as problem types, solution styles, effects of internal and external conditions, and solution processes. Lucas (1972) included a comprehensive survey and discussion of problem solving research and theory in his study. Only the general progress and significant contributions are discussed here.

Past Problem Solving Research

Despite the early interest in problem solving, studies prior to the 1950's were sporadic and progress was slow. Lucas summarized the reasons for this lag in Chapter I; there were semantic ambiguities, overgeneralizations and a lack of consolidation of research efforts. (Green, 1966, p. 3) offered a single explanation. "...there was no major point of view or technique to bring this work (problem solving research) into focus, as Hull's stimulus-response (S-R) theories and Skinner's operant techniques had done for learning." Some helpful directions did result from the early research though. In particular, it was generally agreed that the products of problem solving -- responses, results, or completed methods -- do not permit strong inferences about the processes used. Lucas noted :

> As Bloom and Broder (1950) observe, the process might be inferred from the product if it were possible to design a problem which evokes a unique method of attack. To the writer's knowledge, no such problems exist. Even if such were possible, there are more

> degrees of freedom involved than variation of methods; for example, there are the varieties of heuristics and combinations thereof which could be responsible for production of the same method. (Lucas, 1972, p. 37)

Furthermore, the study of conditions for problem solving revealed little about how people actually solve problems. In order to study problem solving behavior profitably, it was necessary to study subject's overt behavior. Several procedures of varying degrees of utility and reliability were devised to generate and record observable sequences of a subject's behavior. Bourne and Battig (1966) described a sample of frequently employed methods and commented on the limitations. For example, manipulative devices as pendulum problems (Maier, 1931) or jars of water (Luchins and Luchins, 1950) only revealed a few of the hypotheses or hunches a subject was entertaining at a given moment. The limitations of attempting to infer process from external actions made the direct exploration of mental processes a logical alternative.

The direct investigation of problem solving processes required subjects to verbalize their procedures during or after the solution search. Introspection had a subject solve problems and report on his thoughts, reactions, and feelings as he performed. Though introspection externalized thought patterns, there were serious questions about the distortion and interference introduced as a subject analyzed his thinking while solving the problem. Retrospection required the

subject to give a narrative account of his thoughts and processes after he completed this task. Bloom and Broder (1950) found that when problem solving tasks which involved this procedure were used, steps were forgotten and rearrangement of steps into a more logical order resulted. Lazerte's envelope test (1933) explicated a subject's solution path clearly but was not a free choice situation desired in problem solving. In addition to the internal deficiencies described, the two verbalization techniques were expensive in time and equipment, and required careful training by both subjects and observers.

The Thinking Aloud Procedure

One method that avoided some of the difficulties encountered in retrospection or introspection was the thinking aloud technique in which the subject simply verbalized (without analyzing) his thoughts as he worked and his statements were recorded. Kilpatrick (1967) credited Claparede (1917) with originating the technique, and in another publication (Kilpatrick, 1969), indicated that Duncker (1945) was one of the first psychologists to apply this method to mathematical problems.

The thinking aloud technique gained popularity with the application of information processing approaches to the study of problem solving. Newell, Simon, and Shaw (1959) best illustrated such an application in their creation of the General Problem Solver, a computer program that simulates human problem

solving protocols. Green (1966) credited the information processing model with providing a new theory of problem solving and as being the impetus for renewed research efforts. Other researchers have since investigated problem solving, often adapting the thinking aloud technique to their studies.

The thinking aloud method has been subject to criticism and question, but evidence concerning the problem of speech and thinking complementing or interfering with each other has been inconclusive. Kilpatrick (1967) was willing to risk these possible dangers in return for the helpful information that can be gained:

> The method of thinking aloud has the special virtues of being both productive and easy to use. If the subject understands what is wanted -- that he is not only to solve the problem but also to tell how he goes about finding a solution -- and if the method is used with the awareness of its limitations, then one can obtain detailed information about thought processes. (Kilpatrick, 1967, p. 8)

The devices and procedures which were proposed to obtain and maintain observable behavior of subjects in problem solving situations all had limitations. One common assumption was clear though; each technique provided some insight into behaviors related to problem solving. No tests of validity were available for the various techniques, perhaps because of the semantic ambiguities and isolated task situations cited by Lucas in Chapter I. However, the increasing recognition and use of the thinking aloud procedure in research studies provided sufficient reason to assume that the procedure was a

valid one for identifying problem solving behaviors. The patterns and processes revealed by subjects' responses during the interviews added to the investigator's confidence that the thinking aloud procedure actually reflected problem solving behaviors.

Research in general problem solving provided one essential element for this study: the thinking aloud procedure was assumed to be the most valid method for obtaining a direct record of problem solving behaviors. The contributions and directives resulting from problem solving studies in mathematics are examined next.

Mathematical Problem Solving

Kilpatrick (1969), Suydam (1967), Suydam and Weaver (1971, 1972), and Riedesel (1969) provided over two hundred references of mathematical problem solving studies and reports. The numerous citations covered a wide variety of problem solving conditions, problem variables, and problem solver characteristics. Only those studies which are related to mathematical problem variables and to the thinking aloud procedure are reviewed and reported here.

Problem Variables

Riedesel (1969) offered suggestions as he summarized mathematical problem solving research. Three of his ideas involved important factors for the interview test problems: computation, reading comprehension, and item difficulty. He stated, "While the improvement of

computation is important to problem solving ability, the improvement of computation alone has little, if any, measureable effect upon reasoning and problem solving" (1969, p. 54). He also suggested that pupils' reading abilities and item difficulty be considered when designing problems. Though many investigations have studied the obvious relationship of computation and reading skills to problem solving ability, the relative contribution of these skills is not clear. Martin (1963) found a correlation (about .5) between reading and problem solving with computation held constant and a correlation (about .4) between computation and problem solving with reading held constant. Thompson (1967) reported that the effects of reading ability and mental ability on problem solving performances were interactive and that ease of reading was associated with higher performances at both high and low levels of mental ability. Aiken (1972) found inconsistent correlations of language factors and mathematical problem solving achievement as he summarized research findings. Jerman (1971) unexpectedly found that computation was an important factor in his problem solving pilot study and suggested that care be taken that school subjects have adequate practice in basic skills before attempting to solve problems.

Research has not established the interaction of reading and computational skills with problem solving achievement and it appears that the relationship is a complex one. Two logical conclusions are

obvious though. First, written mathematical problems must contain an appropriate level of vocabulary and sentence structure. If a student cannot read a written problem or interpret the meanings of the words in it, he has no chance to solve the problem. Second, the prerequisite computational skills and problem complexity must be appropriate for the mental maturity of the subject. These two precautions were incorporated as the mathematics problems for the interview test and the items for the paper and pencil test were written.

Coding Schemes

In addition to guidelines on item difficulty and subject readiness, research has provided the link that makes the thinking aloud procedure applicable to mathematical problem solving investigations. The verbalized data resulting from this procedure had to be classified. Researchers who used the thinking aloud method to gather evidence also utilized a coding scheme to categorize the recorded behaviors. Gray (1964) used a coding system to classify third, fourth, and fifth graders' solution processes when working multiplication exercises. Bloom and Broder (1950) devised a checklist to code college students' problem solving processes. However, the type of coding scheme used depended on the purpose of the study and a widely applicable coding scheme was not available until Kilpatrick's (1967) study. During his investigation of eighth graders' problem solving achievement,

Kilpatrick produced a general guide for coding protocols audiotaped in word problem solving sessions in mathematics. He restated the thirty-six questions from Polya's How To Solve It (1957) and arranged them on a form so that a tally could be made each time a subject asked himself one of the questions or acted as though he had put such a question to himself. The complexity of applying the original checklist led Kilpatrick to modify the coding procedure, but he was disappointed with the information which was wasted and sequences which were omitted when coding with only a checklist. Consequently, Kilpatrick devised a comprehensive system which included both a checklist and a model for coding the chain of behaviors occurring in a subject's protocol. After applying the system to his subjects, he revised it into a useful classification for identifying individual differences and styles in problem solving behaviors (Appendix A).

Lucas (1972) extended Kilpatrick's classification system in a study involving heuristic problem solving strategies in calculus. During a pilot study, Lucas altered the checklist by splitting the deduction process into analysis and synthesis and adding symbols to represent drawing of diagrams and algorithmic processes. He also made numerous revisions in the process coding system, including elimination, refinement, and addition of categories. Lucas introduced parentheses ( ) to indicate the scope of a production process

and developed a scoring system based on performance within a problem. The reliabilities of both the coding and scoring systems were established by measuring agreement with a second coder. Lucas used his revised coding system to detect changes in heuristic processes as a result of diagnostic teaching of problem solving strategies in calculus, but the form (Appendix B) is easily adaptable to any study involving mathematical problem solving in thinking aloud interviews.

Studies in general problem solving and in mathematical problem solving have combined to produce a reliable and valid procedure for recording and assessing problem solving behaviors. The thinking aloud procedure and Lucas' refined coding system were utilized to record, assess, and rank the mathematical problem solving performances of seventh graders. A second part of this study attempted to develop a written instrument which has concurrent validity based upon the thinking aloud procedure. Some background on test theory is necessary to explain the validity of the written instrument.

Test Criteria and Construction

The thinking aloud procedure and related coding system are assumed to be a reliable and valid means for assessing mathematical problem solving performances. However, the physical and financial implications of the procedures restrict their usefulness. This study explores the feasibility of producing a written test as a substitute

for the expensive, time-consuming procedures of recording, coding, assessing, and ranking pupils' problem solving performances. The procedure of developing another test which gives information similar to that derived by an existing instrument leads to a question of concurrent validity. The different type of validity are featured as related literature on test construction is reviewed.

Test Reliability

One important quality of a good test is reliability. Given a person possessing a certain characteristic of magnitude m, the score observed with some measuring device will be m + e where e is the measurement error. A reliable instrument minimizes the error to some acceptable bounds. Though the error in measurement may be due to external conditions or to the instrument itself, reliability decisions are usually based on variance in measurements of the device. Some recognized statistical tests of reliability are the Kuder-Richardson Formula 20 (1937), Cronbach's alpha (1970), and Hoyt's Measure of Internal Consistencey (1941). The latter will be used as a reliability measure of the written test being designed in this study.

Test Validity

Cronbach (1970) uses the terms "criterion oriented" or "predictive," "content," and "construct" in discussing types of validity.

Construct validity does not apply to this study, but the other types of validity are vital in the two tests being constructed: the interview test (IT) and the written test (WT).

Content validity is essential for the IT. The entire study is based on the definition of mathematical problems and the assumption that the behaviors exhibited by subjects solving these items are mathematical problem solving behaviors. For the test to have content validity, the test items have been judged as having met the criteria of the definition formulated in Chapter I and must be acceptable by the considerations suggested by the related research discussed earlier in this chapter. Content validity is further established by examining items and comparing them to the universe which the test is designed to represent. The content validation procedures of the IT are outlined in Chapter IV. The validation necessary to establish the theoretical basis for the WT is discussed here.

When a person's future performance is predicted from a test score and the expectations are matched against some follow-up criterion, the accuracy of the predictions is the test's predictive validity. The predictive ability between two measures where no time is assumed to have lapsed between them (i.e., the information was obtained concurrently) is concurrent validity. The thinking aloud procedure and related coding scheme are the accepted procedures for gathering data and establishing ranks of students' problem solving performances.

BEST COPY AVAILABLE

The WT is a proposed substitute for the interview and coding procedure and the feasibility of using it as a replacement depends on concurrent validity. That validity will exist if the ranking resulting from the WT correlates well with the one which was validly established by the elaborate methods.

The usual procedure for establishing whether two tests measure the same thing (i.e., have concurrent validity) is to compute a correlation coefficient between the scores or results. Pearson's product-moment correlation coefficient (Hays, 1963, p. 497) and Spearmen's rank order coefficient (Hays, 1963, p. 642-647) are two popular measurements of agreement between two tests. In case of ties in the rankings, Kendall's tau (τ) (Hays, 1963, p. 652-655) can be used. A coefficient of 1 or -1 indicates that a perfect relationship between the two scores exists and that either score can be predicted perfectly from the other. A coefficient of zero means knowing one test score is no advantage in predicting the score on the other instrument. If the correlation coefficient between the results at the WT and the IT exceeded .7, the WT would be judged a feasible alternative to the interview and coding procedure for assessing the mathematical problem solving achievement of seventh graders. As evidenced by the examples below, the use of concurrently valid alternatives to testing instruments and procedures is a common and accepted practice.

Present testing programs rely on concurrent validity, especially for the purpose of creating short forms of long tests. Most mental and achievement test batteries are designed to be administered over lengthy periods of time. For example, the Science Research Associates (SRA, 1964) battery for grades 5 through 9 requires six and one-half hours of actual test writing time. The California Achievement Test (CAT. 1970) battery consists of 381 items in 18 subtests under 3 general headings. It takes two hours to complete, plus additional time for respites and directions between subtests. The time required for either of these two batteries may be prohibitive for individual teachers or schools with crowded schedules. The California Short Form Test of Mental Maturity (1963) is a substitute for the lengthy CAT battery. It has 64 items, requires 34 minutes to complete, and according to Cronbach (1970, p. 138) has a concurrent validity coefficient of 0.77 for the total test score. Obviously, since there is not a perfect correlation, some error is introduced as interpretations are made on the short form test instead of using the complete battery. Decisions about the error tolerance have to be made before deciding whether to use a substitute test, but the great savings in time have made the use of concurrently valid alternatives a common and accepted practice.

Validity of Available Procedures

The proposed written test of this study was similar to other

written instruments which were available commercially or which have been used in previous studies. However, this investigation conceived that there were few existing methods for assessing problem solving achievement and that the validity of these methods was questionable. The examination of prior research verifies the dearth of problem solving assessment instruments. Only the validation procedures of available instruments need be examined to complete the verification. The review now centers on commercial tests and their validity.

A search of both the technical and teacher's manuals of the Iowa Test of Basic Skills (Lindquist and Hieronymus, 1964) battery failed to uncover any validation procedures for their "problem solving" test (A-2). The test had 32 items which included 5 applications of fractions and 18 money situations. Nineteen items required only one step to achieve the solution. No definition was offered with which to judge content validity, and no concurrent measures were mentioned. A study involving the predictive validity of composite scores from an earlier form of ITBS was discussed in the technical manual, but subscores were not used. The teacher's manual contained a table summarizing the mathematical category and skills represented by each item of A-2. However, no interpretation of a subject's score on the test or on individual items was offered. The writers' statement, "The most valid achievement test for your school is that which in itself defines most adequately your objectives of instruction," seemed to summarize

their attitude toward test validity, especially in the area of mathematical problem solving.

The Metropolitan Achievement Test (MAT, 1970) included 35 items under the heading "Mathematics Problem Solving." The items represented an attempt to require more than one obvious algorithm to achieve the answer. For example, item 20 in the advanced battery stated, "Bricks are sold 36 to a box. It takes two boxes to build one fireplace. How many bricks are needed to build 7 fireplaces?" Despite the MAT's improved attempt to include problem solving behaviors, the tests had limitations. First, the items gave five choices instead of being open-ended. Second, one of the choices was always "DK" for "Don't know" and, in 29 of the 35 items, another choice was "Not given" or "More information needed." The limited selection of alternatives encouraged elimination of and checking the given choices, behaviors which are not typical at the beginning of most mathematical problem solving situations.

The MAT manual discussed test validity, but failed to provide a definition of mathematical problems or any interpretation of test results in terms of problem solving skills. The writers stated that content validity was established by examining textbooks, study guides, and mathematics curriculum recommendations. The teacher's handbook offered advice similar to that of ITBS, "Since each school has its own curriculum, the content validity of Metropolitan Achievement Tests must be evaluated by each school" (Durst et al., 1971, p. 32).

BEST COPY AVAILABLE

Construct validity was concerned with "the completeness of the test as a well rounded or representative sample of the content we are hoping to measure, and also the appropriateness of the types used" (Durst, et al., 1971, p. 32). The test writers felt that concurrent validity and predictive validity had little or no meaning as applied to specific tests within achievement batteries and no validity measurements were offered. Split-half reliability coefficients which averaged about .90 across grades were offered to indicate "that the validity of this test is supported by dependability in the test results" (Durst, et al., 1971, p. 42).

The California Achievement test battery used the word "problems" when giving students directions in six out of the 10 test sections in arithmetic although only section D in part 3 of the grade 7-9 battery was explicitly entitled "Problems." The "Problems" test allowed 13 minutes to solve 15 verbal applications of money, averages, area, volume, and percents. Eight of the "problems" required only one operation while seven items required two operations. Content validity of the test was discussed briefly: it was based on widely accepted mathematics curriculum objectives in the United States.

A perusal of available intelligence and achievement tests revealed two general practices. First, test validity was usually content oriented as the items were chosen to be representative of widely recognized curriculum objectives. This practice meant that the choices

for mathematical "problems" were usually verbal applications (often referred to as word or story problems) of money, percents, measurement, and other familiar topics. Secondly, if a test did not include a "Problems" category in its mathematics section, then an alternate such as "Applications" or "Reasoning" was used. The items in all three categories were essentially the same and no special validity measures were offered except the usual content validity argument.

Commercial tests which claimed to be content valid offered no definition to judge their "problems" by, nor did they establish any relationship between test scores and problem solving behaviors. Thus, the examination of commercial tests partially substantiated the speculation that there were few instruments available for validly measuring mathematical problem solving achievement. Only the validity of available research procedures for assessing mathematical problem solving achievement remained to be examined. Since this study was attempting to produce a written instrument which has concurrent validity with the results of the thinking aloud and coding procedures, only the validation procedure of investigations which used similar procedures or shared a similar purpose were checked.

Bloom and Broder (1950) examined variations in problem solving characteristics of college students through the thinking aloud procedure. They analyzed and scored the subjects' performances, but the investigators did not use mathematical items and the scoring

system was based on the number correct answers chosen from a list of alternatives. Bloom and Broder offered no definition with which to judge validity, nor did they offer any variation measurements or procedures.

Kilpatrick (1967) did not establish any definition, but his "problems" indicated that he was interested in items which necessitated procedures more complex than simple recall or application. His criteria for the "Mathematical Problems Test" (Kilpatrick, 1967, p. 38) agreed with those established in this study and two of his items were used in the pool of problems for this study. Kilpatrick administered a twelve item problem solving inventory to eighth graders in a thinking aloud interview. He analyzed the solutions of items through his coding system and compared the characteristics he observed to results on other tests the subjects had taken. The relationships he noticed are interesting, but the validity of his procedures is the focus here. Kilpatrick used the thinking aloud procedure to obtain a direct audio taped record of subjects' problem solving processes. He based his coding scheme upon thirty-six questions found in Polya's heuristic scheme (Polya, 1957). No validity measurements were made as Kilpatrick logically assumed that he was observing and analyzing mathematical problem solving behaviors.

Lucas (1972) adapted Kilpatrick's coding scheme to study the effects of heuristic training on calculus students. His modifications were discussed earlier in this chapter. Lucas did not measure or question the validity of Kilpatrick's procedures and evidently assumed that he was observing genuine problem solving behaviors. However, the adjustments Lucas made to the coding system were based on direct observation of behaviors in subjects' protocols.

Lucas established a definition of the mathematical problems to be used in his study, but he did not indicate that a pool of items had been formed. Presumably he formed only as many items as were necessary for the interviews. The restricted sample of items could cause questions to be raised about the validity and generalizability of his observations, but this restriction does not undermine the validity of the thinking aloud procedure and coding system for recording and assessing mathematical problem solving achievement.

The validities of commercial instruments and research procedures were examined and the findings supported the hypothesis about the lack of valid problem solving assessment tools. The available commercial tests have inappropriate items and are not supported by problem solving validation procedures. The studies of Kilpatrick and Lucas provided the conception of mathematical problems which was incorporated into this study. They also used recognized

BEST COPY AVAILABLE

procedures for validly assessing problem solving achievement. The investigator is not aware of other studies which attempted to develop a written test as described in this study. Thus, the contributions which have resulted from the examination of general and mathematical problem solving research, of concurrent validity practices, and of available mathematical problem solving assessment procedures are summarized to conclude this chapter.

Summary of Chapter III

General research developed the thinking aloud procedure as a valid means of obtaining direct records of subjects' problem solving processes and the procedure was made amenable to assessment of mathematical problem solving behaviors through the efforts of Kilpatrick and Lucas. The protocols resulting from thinking aloud interviews can be analyzed through a coding scheme to provide a valid assessment of mathematical problem solving. Other procedures claimed to assess mathematical problem solving, but an examination of them revealed serious limitations which cast doubt on their validity.

Previous studies provided guidelines which were used to create an improved written instrument and statistical practices established procedures which were to judge its validity. The investigator established a definition of mathematical problems and controlled the reading and computation difficulty of the items. Accepted procedures

for establishing concurrent validity were followed. Chapter IV includes the detailed steps which were followed as it traces the design of the study.

BEST COPY AVAILABLE

Chapter IV

DESIGN OF THE STUDY

Introduction

There were three principal parts to this study. First, the problem solving performances of students who were observed in interview situations were recorded, analyzed, and ranked. Second, a proposed written test (WT) was devised, developed and administered to the same students to provide a second ranking. Third, the correlation between the two ranks was determined and necessary adjustments were made in the procedures of the first two parts. The detailed plans for each part of the study are discussed separately.

Part I: The Complex Problem Solving Assessment Procedure

A recognized valid procedure for measuring a subject's mathematical problem solving achievement is to interview him in a thinking aloud situation and code his reponses. In this part, the plans for utilizing the complex assessment tool are developed. Consideration was given to the mathematical problems, the subjects, the interviews, the coding system, and the ranking procedures. The interview test items which formed the basis for the measurement of mathematical problem solving achievement are discussed first.

The Mathematical Problems

The interview test (IT) problems were to be drawn from a pool which was to be developed by the investigator according to the definition of "mathematical problem" given in Chapter I. (See Appendix C for the complete IT item pool). The judgments of authorities about the items, the results of a pilot tryout of the items, and an examination of the mathematics curriculum as described by textbooks were to be used to screen items and strengthen content validity. Items which did not receive a majority of the judges' approval were to be rejected or to be rewritten and resubmitted. Mathematical problems which pilot study subjects found difficult to read or to understand were to be rejected. Items which included content too advanced for most seventh graders (e.g., solving quadratic equations) were to be revised or omitted. In addition, the size and kind of numbers in the problems were to be simple enough to prevent computational difficulty from being an important factor.

It should be noted that the problems for the IT were not to involve the use of special apparatus. The only materials available would be paper and pencil as Lucas' and Kilpatrick's coding schemes were designed for the oral and written responses to mathematical problems which were presented in writing. Subjects working with special devices or materials might perform manipulations which are

BEST COPY AVAILABLE

visually apparent, but not verbalized for recording on tape. An observer could note these physical actions, but would need a new coding scheme and might not accurately record a sequence of actions. The use of video tapes and the development of alternative coding schemes are legitimate areas for future research, but were not included in the original plans for the study.

The Subjects

In Chapter II, it was indicated that the IT and WT were designated for the seventh-grade level. Thus, the items for both tests were to be chosen to represent this level of mathematics, and the subjects were to be seventh graders. The restriction to seventh grade was chosen for several reasons, but two factors had the most influence. First, it was necessary to restrict the scope of the study. If several grades were to be tested, then a separate item pool would have to be established for each test at each level and subjects at each level would have to be interviewed. Time alone would be a deterrent to such an undertaking.

The second reason for the restriction to seventh grade was the interview and coding procedure itself. Kilpatrick (1967) used the thinking aloud procedure with eighth graders, but only chose subjects with above average mathematical ability in order to avoid undue frustration. His coding system was developed from his observations of the eighth graders. The investigator became interested in mathematical

problem solving during a state assessment of seventh graders and wanted to continue with this level. It was assumed that Kilpatrick's coding scheme would be applicable to seventh graders because of the close age proximity and the natural grouping of seventh and eighth grades in almost every curriculum design. Kilpatrick's precaution to avoid undue pressure or frustration was incorporated into this study: only seventh graders who were classified as average or above average in mathematics achievement would be eligible for the interviews. It was planned that available achievement test scores would be used to separate subjects. If no tests were available, a standardized test such as the Stanford Achievement Test was to be administered to screen the potential subjects.

The Interviews

This portion of the study utilized the efforts of Kilpatrick (1967) and Lucas (1972). Seventh graders were to be asked to solve six mathematical problems in a thinking aloud interview and their statements were to be audio-taped. The procedures for conducting the interviews were well developed and explicit. Subjects would be asked to sit at a table with an observer sitting close enough to notice their writing and reactions as they solved the problems. A few minutes would be taken at the beginning to put the subject at ease, and instructions (Appendix E) would be read with the subject before he attempted to solve any problems. One or two sample problems

BEST COPY AVAILABLE

would give the subject a chance to adjust to the interview situation and would enable the observer to comment on the subject's thinking aloud habits.

During the interview, there would be two possible sources of concern: (1) the subject's inability to solve a problem which would lead to (2) the time factor. The former was circumvented through the directions -- the student could leave the problem if he had attempted to solve it and was repeatedly thwarted or if he stated that he could make no further progress. This option should have prevented an excessive amount of time being spent on a single problem, but an additional precaution was necessary to avoid a tediously long session. If the subject did not have an opportunity to solve each of the problems in an hour or if he appeared to be tiring, it could have been necessary to have him return at another time. Such a decision would be made by the observer.

The interview rules for the subject were contained in the instructions. He was to read the complete problem aloud before attempting to solve it and was asked to think aloud as he was seeking a solution. He would be permitted to rest after completing a problem or after requesting to leave a problem which he was unable to solve.

While the interview session was in progress, the observer would control the audio-taping equipment (tape-recorder) and make notes on the subject's progress. He could answer a question if it would not assist the subject in solving the problem. He was not to interfere

with the subject's progress unless the subject fell silent for a set period of time. Lucas asked a question like "What are you thinking now?" to prod a subject who had neither spoken nor written for a period of 30 seconds. His approach was to be followed in this study, and his practice of not discussing a problem or informing a subject of the correctness of an answer was also included. The interview would be concluded after the subject had read and attempted to solve the last interview problem. He would not be permitted to return to any problems which he left unsolved nor could he go back over those he completed.

After the interviews were conducted and audio-taped, the resulting protocols would be coded, analyzed, scored, and ranked. Again, the procedures of Kilpatrick and Lucas were to be used.

The Coding System

The coding scheme for this study was a combination of Lucas' and Kilpatrick's. The former is a modification of the latter, but since Lucas designed his for use with calculus students there were items such as "implicit differentiation" which had to be eliminated. Other revisions would be made according to the results of a pilot study (discussed in Part III). Categories or items which were unique to an individual would be discarded as they would not provide information about problem solving patterns. Items which yielded poor interjudge reliability ($< .5$) would be examined closely for revision or rejection. Finally, the unnecessary coding symbols such as

BEST COPY AVAILABLE

Me_{jdiffi} = model introduced by equation j obtained from i by differentiation" would be eliminated. Possible alternatives to elaborate symbols such as "$\ddagger\ddagger$" were also to be considered.

Lucas developed a five point scoring system based on a subject's complete protocol for a problem. Points were divided into three categories: Approach, Plan, and Result. "Approach" represented the subject's understanding of the problem and was evidenced by correct interpretation of the data, conditions, and objective. One point was awarded if the subject avoided or nullified structural errors due to misinterpretation. A subject's "Plan represented the relationships and solution path which he developed in attempting to achieve the answer. A maximum of two points was awarded when if all structural errors were corrected or nullified and it was clear that the problem could be solved correctly in the absence of executive errors. The subject's "Result" was his final answer and a maximum of two points was awarded if the solution was in a correct form. The complete scoring scheme was discussed by Lucas (1972, p. 177-181) and is summarized in Appendix F of this study. His scoring procedure was followed as the ranking of students was developed.

The Ranking

There are many ways to rank subjects, but only the direct results of their interview performances should be used if a valid measure of mathematical problem solving achievement is desired. Two measures

would be available after Lucas' coding and scoring systems were applied to subjects' protocols. The number of items correct would be the simplest measure. The total process score (or any of the subscores) would provide another basis for ranking subjects. Both the number correct and the process score were to be considered in order to determine a correlation of students' interview performances with their ranking on the written test.

Another basis for ranking interview subjects could be derived by statistical analysis of their protocols. Latent partitioning (Torgerson, 1958, or Lord & Novick, 1968) or a type of clustering analysis (Hubert, 1973) was to be applied to the coded data. Based on patterns of the coded behaviors, these statistical procedures should provide a separation of the subjects into subgroups. Then an ordering (to be determined by the investigator) between and within the subgroups should provide another ranking to compare to the results of the WT.

Summary of Part I

The components of Part I have been identified and detailed. It was noted that the interview and coding scheme were to follow the established procedures of Kilpatrick and Lucas. The ranking and analysis procedures were exploratory since they were subject to adjustments depending upon the strength of the correlations computed in Part III. The IT item pool was to be established through usual validating procedures. The combination of these components should have produced a ranking which was assumed to be a valid measure of the

BEST COPY AVAILABLE

subjects' mathematical problem solving achievement. Part II describes the development of the proposed written test (WT).

Part II: The Written Test

The purpose of this part of the study was to devise a paper and pencil instrument (WT) which would provide a second ranking of the same subjects who participated in the interview procedure of Part I. Subjects were to take the WT and to be ranked according to the results. Hopefully, the statistics of Part III would yield a high correlation between the rankings of Parts I and II. A high correlation would provide the concurrent validity necessary to establish the feasibility of using the written test as a substitute for the complex interview and coding procedure.

The WT Items

In addition to a high statistical correlation, it was also desirable that the WT have a logical relationship to mathematical problem solving. These considerations helped determine the choice of items for the paper and pencil instrument:

1) The items were to be mathematical in nature, covering topics as arithmetic, algebra, geometry, probability, and logic.
2) The items were to be non-routine in order to promote some of the same reasoning processes involved in the mathematical problems of Part I.
3) The items were to be open-ended. Choices of answers would not be provided.

4) Any numerical manipulations or symbols were to be within the ability level of most seventh graders, (i.e., square roots and the corresponding notation could be unfamiliar to them.) It was assumed that the four basic operations and combinations of them could be handled by seventh graders.

5) The size of the numbers involved was to be small so that it would not promote computational errors.

6) The number of steps involved should consider the time available and the ability and maturity of seventh graders.

For example, a combination problem could require the subject to find twenty possible solutions which might take more time and patience than the subject wished to spend on it even though he could ultimately arrive at the solution. Three mathematical operations could be a logical maximum for a direct solution as in this example: The perimeter (distance around) of a rectangle is 46 feet. If the length is 15 feet, what is the width of the rectangle? The direct solution includes multiplying 15 by 2 (or adding 15 and 15), subtracting 30 from 46, and dividing the difference by 2.

It should be noted that the WT items were not the same as the mathematical problems used in the IT. The items in the WT could require one step solutions and they need not meet the criteria of

mathematical problems, but would attempt to avoid simple recall of knowledge or direct applications of algorithms.

A second item pool was to be created according to the criteria listed above. Again, interjudge agreement, pilot study results, and an examination of the mathematics textbooks were to be used to screen items. The completed pool (Appendix D) was used to provide a sixteen item random sample for the written test.

Administration of the WT

In order to consider a school situation, the WT would be devised so that it could be administered to one or more students in one class period of approximately 50 minutes. Directions to the student would be provided on the first page. The administrator would read the directions aloud with the subjects and answer questions before the test starts. At a signal from the tester, the pupils would begin attempting to answer the questions and were to put their solutions in the spaces provided or on the diagrams included in the items.

In this study, the students were to write directly on the test since adequate space would be provided between the items.

As the subjects were taking the WT, they would be permitted to skip items and return to them later if they desired. When a student completed or attempted all the items, he was to return the test to the administrator.

The test administrator was to hand out the tests, read the directions, and answer students questions about the directions. During the test, the administrator would need to watch for students copying answers from others. He could answer students' questions if the information would not assist the student in finding the answer. The tester could not provide definitions to words in the items, explain the question in an item, or provide hints about solutions or solution procedures. The test administrator was to collect the tests as the students completed them. By the end of the period it was assumed that all students would have completed the items. If some students were still working, they should have been permitted additional time. This condition was crucial if the subject had a number of items to complete! The written test was designed to be a power test and it was assumed that all subjects would have as much time as they need. Common sense and the discretion of the test administrator would have prevailed if a subject were still working after one hour. If it appeared that the student was making little progress or if he did not have many items left, it would probably not seriously affect the rankings if he were asked to turn in the test.

The Ranking

On the WT, the subjects would be ranked solely on the basis of their correct responses. Three factors influenced this decision:

BEST COPY AVAILABLE

computational difficulty and reading level of the problems, and a process-product consideration.

Jerman (1971, p. 81-83) encountered unexpected complications in his study when he used only correct solutions as his criterion measure: computational difficulty was not considered in his pilot study, but it emerged as an important factor in his experiment. His reaction suggested that if he had delayed his study, there would have been more time for teachers to thoroughly review basic skills with a result that the students would have made fewer computational errors on the test.

Jerman's experience provided two precautions which were considered in this study: 1) since this study involved seventh graders about midway through the school year, it was assumed that the subjects would have learned and reviewed the basic skills necessary to do the computations on the WT and 2) the difficulty level of the items would be controlled so that the computations involved would be familiar to most seventh graders. The investigator assumed that these two precautions would have prevented computational difficulty from becoming an important WT factor. Furthermore, the WT differed from Jerman's written tests because the instrument devised for this study would be focusing on the rank of the students and the investigator assumed that computational errors would not affect the ranks of the students. In practice, it was assumed that the ranks on the

written test would remain the same if credit were given for answers which would have been correct if a computational error had not been made.

The relationship of reading ability to problem solving ability is uncertain. However, if a student cannot read a written problem or interpret the meanings of the words in it, he has no chance to solve the problem by himself. Another precaution was necessary: the problems on the written test were to be expressed in a language simple enough to be read and interpreted by most seventh graders. However, it is possible that despite the given precautions, some subjects who have been included in the study might not be able to handle either the mathematics or the vocabulary of the problem. Such difficulties could not be noted on the WT, but the reading level of the items was to be kept simple enough so that the investigator could assume that this factor would not affect the rank of the students.

Researchers agree that the final product (the solution) to a problem does not reveal the processes involved in arriving at the result. Jerman (1971, p. 74-76) exemplified this contention when further analysis based on the correctness of procedure revealed significant differences which were not apparent when only correct answers were used as the criterion. However, the WT in this study was not designed to measure the processes involved in arriving at

BEST COPY AVAILABLE

the solution. The sole purpose of the WT was to provide a second ranking of the same students who were given the IT. The possibility of giving partial credit based on procedures or on incomplete answers would have necessitated the establishment of special instructions to students, detailed observation procedures of students' work, a new rating scheme, and assumptions which were beyond the scope of this study. In addition, these numerous extensions would have complicated the WT beyond the simplified form that could be used for large scale testing.

Summary of Part II

The planned development of the WT has been described. In summary, a 16 item sample of the WT item pool would be presented to the same subjects who participated in the interviews. The subjects were to be ranked according to the number of items they answered correctly and the ranking was to be compared to the one resulting from the IT of Part I. Revisions in the WT were to be made according to the results of a pilot study and the statistical comparison of the rankings. The statistics are described next.

Part III: The Comparison of Ranks

Part III was designed to seek and test similarities between the rankings developed in Parts I and II. One direct test of a relationship between the ranks was the size of a correlation coefficient. Other similarities would be sought by examining subjects'

protocols and utilizing statistical procedures to seek problem solving patterns. Steps which were used in Parts I and II could also be examined for possible changes if there was a low correlation between ranks.

Correlations

After the rankings from Parts I and II were established, statistical procedures would be applied to search for a correlation between the ranks. Spearman's rank order correlation coefficient was to be computed. In case of ties in either ranking, Kendall's τ (tau) (1955) could be used to measure the agreement between the ranks. A correlation of at least .71 would indicate that the WT scores account for approximately fifty percent of the variance in the IT ranks. It was decided by the investigator that this result would be the minimum correlation to support the feasibility of using the WT as a substitute for the complex thinking aloud and coding procedure. A correlation below .71 would lead to a number of questions in all three parts of this study. The procedures, codes, analysis, and scoring of the interview-coding system would be examined in search of possible adjustments to increase the correlation. The items, length, and scoring of the written test could be altered, or the statistical analysis itself might be changed. Especially suspect would be statistical analyses of the protocols of the subjects in Part I. Their experimental nature requires additional discussion.

BEST COPY AVAILABLE

Processes and Patterns

The procedures for computing correlation coefficients are well established and would not be altered in the search for a stronger relationship between the ranks of Parts I and II. However, the exploratory ranking procedures mentioned in Part I could be altered in search of improved correlations. The application of latent partitioning or clustering analysis to problem solving protocols was an experimental step in this study. Ideally, either procedure would provide a separation of subjects into subgroups with one or more member in each. Then the subgroups would be ordered to provide a ranking of subjects. If there was more than one subject per subgroup, an ordering between and within subgroups would have to be developed in order to rank the subjects.

Since the subgroupings and orderings could not be determined *a priori*, decisions about the statistical procedures were to be made after the data had been collected and coded. If either latent partitioning or clustering provided a distinctly advantageous subgrouping, then the rankings based on that procedure would be retained. If both procedures provided similar subgroupings, then each ones results would be examined during the search for an improved correlation between the ranks of Parts I and II.

Testing Procedures

Any testing program demands certain precautions and controls. The checks for validity and reliability have been discussed. Other

procedures were necessary to prevent contamination or undesirable factors from entering the study. These were the precautions which the investigator planned to follow:

1) To control for learning effect between the two tests, the subjects were to be randomly divided so that one half would take the WT before the IT while the other half would take the WT after the IT.
2) To avoid excessive fatigue, disinterest, or other negative reaction, the subjects would not participate in both parts of the study on the same day.
3) To avoid a learning effect between items, the problems on both the IT and the WT were to be presented in a random order to each student.
4) Since it was possible that other factors besides problem solving ability could affect a student's ranking in either test, additional information was to be collected on the students. As described previously, students who did not score at the seventh grade level in mathematics achievement would be eliminated from the study. Mathematics achievement could also be examined for relationships to response patterns and tendencies. Students who could not read and interpret at least five of the six IT problems would not

provide sufficient problem solving data to be included in the study.

5) The effect of participating in the study could influence a subject's problem solving performance. It was assumed that this effect would affect both the IT and the WT rankings similarly and thus not significantly affect the correlation between them.

A pilot study with about ten students would be conducted to give practice in using the interview and coding procedure. It would also help refine the available instruments and guide the selection or rejection of problems to be used. The refined system was to be applied to a larger population of thirty to forty seventh-grade students.

Interpreting Results

Three questions were posed during the discussion of the specific problem in Chapter II. Since this was an exploratory study, the answers to these questions were sought in various ways. Each question is considered separately.

The first question asked "How well does the thinking aloud procedure and related coding scheme capture and classify the mathematical problem solving behaviors of seventh-graders?" The observation of subjects in the thinking aloud interviews could produce empirical evidence about the effectiveness or deficiencies of the thinking

aloud procedures. If the coding scheme did not accommodate the problem solving behaviors of seventh-graders, then adjustments would be suggested.

The second question, "Is it possible to assess, separate, and rank seventh-graders according to their coded mathematical problem solving protocols?", depended on the application of Lucas' scoring system and the results of the exploratory partitioning procedures. If the scoring system or subgrouping procedures made it possible to develop ranking procedures, then the answer would be affirmative. Hopefully, the resulting ranks would have a strong relationship to the written test ranks. A negative answer could result if the scoring system did not permit logical rankings to be developed or if the data from the coded protocols did not fit exploratory partitioning schemes.

The third question was the central focus of the study, "Is it feasible to construct a written evaluative instrument whose results correlate well with the ranking derived from the coded protocols?". The feasibility would be determined by the correlation coefficient: above .71, the written test would provide satisfactory predictive powers; below .71, the WT would account for less than one half of the variance of the ranks of the IT. Feasibility could also be influenced by physical features of the WT. For example, if a satisfactory level of reliability (.84 or more) could only be reached by increasing the length beyond a reasonable time period (one hour),

BEST COPY AVAILABLE

then the feasibility of the WT would be questionable.

The procedures for using the data to answer a priori questions have been discussed. Other questions which might arise during the course of the study would have to be handled individually.

Summary of Chapter IV

The details of Parts I, II, and III of this study have been presented. Part I was designed to produce a valid ranking of subjects by their problem solving protocols. In Part II, a written test would be developed to produce a second ranking of the same subjects who participated in Part I. Part III was the search for a sufficiently high correlation between the ranks of Parts I and II so that concurrent validity was established. The feasibility of using the written test as a substitute for the thinking aloud and coding procedure would be determined by the strength of the concurrent validity. The execution of these plans follows in Chapter V.

BEST COPY AVAILABLE

Chapter V

EXECUTION OF THE PLANS

Introduction

The preceding chapter presented the plans and design of the study. Chapter V reports the execution of the plans and the deviations from predetermined procedures. A pilot study accounted for many of the changes and is reported prior to the description of the main study.

Pilot Study

The purpose of the pilot study was to give the investigator an opportunity to practice conducting interviews, to test his reliability in using Lucas' coding and scoring schemes, and to try a form of the written test (WT). The pilot study proceedings and results were used to suggest changes in the original plans of the study. Modifications were made in the taping format, in the questions which were to be answered during the study, in the WT length, in the interview procedures, and in the checklist and coding scheme. The changes are identified as each part of the pilot study is discussed.

Pilot Study Sample

During the summer of 1973, eight volunteers who had completed

seventh grade in Madison, Wisconsin, took both the WT and IT. Seven subjects had attended Van Hise Middle School while the eighth had attended Cherokee Middle School. In order to avoid a fatigue factor, no subject took both tests on the same day.

The WT's were administered in a community building at Eagle Heights, the University of Wisconsin's married student housing. Five subjects were tested on one occasion while the other three were given the WT individually on separate dates. All of the WT's were administered prior to the IT's.

The IT's were to have been administered during individual sessions at the community building. However, the audio taped interview with the first pilot study subject suggested a change in the type of tape to be used for data recording. As a result of the change, only four subjects were audio taped as planned: the remaining four were video taped. The circumstances leading to the trial of video taping and the results of this change in media are described next.

Audio Taping Versus Video Taping

After audio taping the interview with the first subject, it became apparent that interesting physical actions and important silent indicators of problem solving processes were not captured. For example, a subject could move his pencil across the sentences of the interview problem as he silently reread. His lip and eye movement verified that the subject was indeed rereading. However, the audio

BEST COPY AVAILABLE

tape recorded silence where this important behavior occurred. Therefore, the investigator decided to use video taping on four pilot subjects to explore the advantages of getting a visual and audio record of the interviews. The video taped pilot study interviews were conducted in the Wisconsin Research and Development Center. A later decision incorporated the use of video tape into the main study in order to get more reliable feedback of subjects' problem solving behaviors.

The use of video tape in the pilot study prompted further questions, the first being, "What differences are there in the information gained between audio taping and video taping subjects?". Plans were made to answer this question during the main study. One procedure required that a sample of video taped protocols be coded twice; once with the video and audio together and once without the video, in random order to avoid a coding practice effect. Inter-judge unreliability could account for some differences between the two codings, but if the same coder did both, the judgment differences should be minimal and minor. Obvious differences in the two codings of the same protocol should be apparent by inspection. For example, during the pilot study the video tape clearly indicated when the subject was introducing or changing a diagram whereas the coder could only infer such actions from the audio portion of the tape.

A second plan for seeking differences in the information gained was a statistical comparison between the codings developed from the audio-video and the audio-only playbacks of the same tape. A strong agreement would indicate that the video advantage did not provide much additional information. It was planned that the agreement measure would be made on the process codes, the checklists, and the points awarded. The measure would be the quotient of the number of times the two codings agree and the number of chances for disagreement to arise.

A general comparison of the coded information resulting from audio taping and from video taping was considered as a third way to detect the differences in information gained. However, it appeared that the pilot subjects who were video taped behaved differently than if they had been audio taped: those subjects whose interviews were recorded on video tape appeared to be in a greater hurry to complete the problems and seemed to be more nervous than subjects who were audio taped. Thus, the general comparison of information gained from audio and video taping could not reliably be made.

A second question arose: Do subjects perform differently if they are video taped instead of being audio taped? Two measures of difference based on problem solving scores and time were proposed to answer this question. The problem solving interview scores of the

BEST COPY AVAILABLE

two groups would be compared through a one way fixed effects analysis of variance (ANOVA) with the two groups randomly assigned to treatments (the two taping procedures). There are two logical sources of differences: one is that the presence of distracting and novel video taping equipment may adversely affect the subjects' scores; the other is that the novelty of the situation might motivate the students to perform better than they would in the presence of familiar equipment as a tape recorder. Thus, the following hypothesis was posed:

> Hypothesis H1: The mean score on achievement for video taped subjects equals the mean score on achievement for audio taped subjects.

An arbitrary .05 level of significance was chosen for rejection of the null nypothesis.

It was also planned that a one way fixed effects ANOVA could be applied to the total amount of time each subject used to solve the six mathematical problems given during the interviews. A significant difference in the amount of time used by subjects would encourage questions about the performance differences due to the two taping routines. Thus, a second hypothesis was posed:

> Hypothesis H2: The mean solution time of the video taped subjects equals the mean solution time of the audio taped subjects.

An arbitrary .05 level of significance was chosen for rejection of the hypothesis. Figures 5.1 and 5.2 contain the ANOVA designs for the comparisons of process scores and solution times of audio taped and video taped subjects.

Treatment X (video tape)	Treatment Y (audio tape)	
X_1	Y_1	
X_2	Y_2	H1: $\mu_x = \mu_y$
.	.	
.	.	p<.05 could be significant
.	.	
X_n	Y_n	

X_i = Total process score of video tape subject i

Y_i = Total process score of audio tape subject i

Figure 5.1. One Way Fixed Effects ANOVA of Process Scores

Treatment X (video tape)	Treatment Y (audio tape)	
X_1	Y_1	
X_2	Y_2	
X_3	Y_3	H2: $\mu_x = \mu_y$
.	.	
.	.	p<.10 could be significant
.	.	
X_n	Y_n	

X_i = Total time video taped subject i used to attempt the six interview problems.

Y_i = Total time audio taped subject i used to attempt the six interview problems.

Figure 5.2. One Way Fixed Effects ANOVA of Solution Times

BEST COPY AVAILABLE

The incorporation of video taping into the study evoked one issue which was not directly related to the data. Lucas coded the protocols obtained during the pilot tryout in this study and made an observation that it took noticeably less time to code video taped protocols than to code audio taped protocols. One possible explanation for this difference was that the video tapes made subject actions more apparent, thus requiring a smaller number of replays than would be necessary to make judgments from audio tapes. The data from the pilot study supported Lucas' hypothesis: the audio tapes took approximately two and three fourths (2.7) minutes of coding per minute of recording while the video tapes took slightly less than two (1.9) minutes of coding per minute of tape. The apparent differences in time could be an important factor for an investigator faced with numerous or lengthy interviews. It must be noted that the statistics were gathered on a recoding of the pilot protocols and this repetition introduced other variables such as the effect of coding practice and familiarity with the tapes. A systematic scheme for testing the coding time differences was necessary before more reliable conclusions could be made. Thus, each taping procedure was considered a treatment and the subjects of the main study were to be randomly assigned to permit an ANOVA. The hypothesis to be tested was the following:

> Hypothesis H3: The mean coding time for audio taped protocols equals the mean coding time for video taped protocols.

An arbitrary .10 level of significance was chosen for rejection of the hypothesis.

Because a statistically significant difference in coding times may not be important in practice, a second method of comparing coding times was planned. The difference between the average coding time for one minute of audio tape and the average coding time for one minute of video tape would be found. If the absolute value of the difference represented a 10 percent advantage of one type of tape over the other, the difference could be important. Figure 5.3 contains an outline of the two tests for differences in the coding times.

One Way Fixed Effects ANOVA on Coding Time

Treatment X (video tape)	Treatment Y (audio tape)	
X_1	Y_1	
X_2	Y_2	
.	.	H3: $\mu_x = \mu_y$
.	.	
.	.	p<.10 could be significant
X_n	Y_n	

X_i = Total time to code the tape of video subject i

Y_i = Total time to code the tape of audio subject i

A Direct Comparison of Coding Time Differences

A = Number of minutes to code one minute of video tape
B = Number of minutes to code one minute of audio tape

$|A - B|$ could be important if the difference is at least 10% of A or B

Figure 5.3. Two Ways to Test for Coding Time Differences

The modifications resulting from the incorporation of video taping into the study have been described. The information gained should

BEST COPY AVAILABLE

be helpful to future researchers who consider using video taping as a data recording instrument. It will help answer questions concerning the effectiveness of the thinking aloud procedure. Other changes resulting from the pilot study follow.

Changes in the Written Test

The original written test contained sixteen items. However, Hoyt's internal consistency measure produced a reliability of only 0.1765. (See Appendix H for data). The extremely low reliability could have been due to the small number of subjects in the pilot study, an unusual interaction of subjects and items, or the number of items in the test. It was assumed that the first two possibilities would be compensated for in the main study by the larger sample of subjects and the random item sampling procedure. The third possible cause of low reliability was counteracted by increasing the length of the WT from 16 to 20 items. A higher number of items would strengthen reliability further, but could require more than the desired time maximum of one hour for the students to complete.

The increased length was the only change made in the WT. Modifications which were made in the interview procedures are described next.

Changes in the Interview Procedures

The pilot study produced two changes in the interview strategies. First, the apparent nervousness and haste of pilot subjects who were video taped suggested that extra efforts would have to be made to put students at ease before having them solve the problems aloud.

BEST COPY AVAILABLE

It was planned that during the introductory comments and reading of directions, the observer would verbally emphasize that the interview was not a speed or accuracy test and that the subject could use as much time as necessary. Care was to be taken so that the observer did not appear anxious or hurried (e.g., he would not overtly look at the time or place a timepiece in a conspicuous position). In order to put the subject further at ease, the observer would converse with the subject until it appeared that the student felt comfortable before the cameras. The same precautions were also planned for subjects who would be audio taped although the presence of an ordinary tape recorder did not appear to unsettle any of the pilot subjects.

The second change in the interview procedures resulted from Lucas' practice of verbally encouraging a subject to think aloud if the student fell silent and inactive for a period of thirty seconds. When one pilot subject was prodded with "What are you doing now?" after a silence of thirty seconds, he appeared slightly irritated at having his thoughts interrupted, replied "I'm thinking", and lapsed back into silence. Similar reactions by other subjects persuaded the investigator to avoid interfering with subjects who fell silent for more than thirty seconds in the main study. The discretion of the observer was to be used if the subject became silent and not overtly active for more than a minute especially if it appeared that the subject was stymied or frustrated. But care was to be taken that the observer would not interrupt a subject if it appeared that he was silently devising a plan, even

BEST COPY AVAILABLE

though this neglect would cause gaps in the thinking aloud record of a student's problem solving procedures.

In addition to the changes in the interview procedures, some modifications of Lucas' coding system were suggested by the pilot study. These alterations are discussed next.

Changes in the Coding System and Checklist

Some changes were made in order to adapt Lucas' coding system to the protocols of seventh graders. The subjects in the pilot study never elicited behaviors which were coded as M_{f_c} (introducing diagram with coordinate system imposed), V_s (varies the process), or V_m (varies the problem). These symbols and the related items on the checklist were eliminated from Lucas' format. Another alteration was made when subjects displayed behaviors which were not easily classified by Lucas' system. Additional symbols which were devised to classify the processes were Rr (rereads the problem or parts of it), Rs (restates the problem in his own words), DX (exploratory work with data), TR (random trial and error), and TS (systematic trial and error).

The Rr was introduced because rereading was an important problem solving process and also an indicator of problem difficulty, but Lucas did not have a symbol for it. The Rs was included because pilot subjects rephrased questions as an aid to problem solving. Lucas used DS (deduction by synthesis) as a symbol for processes which combined data to produce new facts or intermediate results, but seventh graders in the pilot study sometimes combined data without

BEST COPY AVAILABLE

any apparent reason or direction. The symbol DX was added to code these undirected deductive processes. For coding trial and error behaviors, Lucas used a T. The pilot study subjects used trial and error often, but displayed differing degrees of sophistication. Some subjects randomly made three or four unrelated guesses before narrowing the range of alternatives or beginning a systematic sequence. The random trials were labeled TR. Other students used the given data to determine their first or second guess. This action resulted in a predictable sequence of trials and was labeled TS.

The above changes in process symbols were accompanied by modifications in the items in the checklist. Appendix G contains the revised process-sequence coding system and checklist. The items which do not appear in the revised list were dropped because no student in the pilot study exhibited such behavior and it was unlikely that any seventh grader in the main study would do so. The other changes are a classification for structural errors and a substitution of letters for symbols when noting errors in the code.

The numerous changes suggested by the pilot study caused considerable delay in executing the main study, but they prevented the investigator from unconsciously increasing stress upon the subjects and from making many last minute changes in the main study.

BEST COPY AVAILABLE

Main Study

The Main Study was conducted according to the modified plans resulting from the pilot study. The population and events of the testing procedures are described after the preliminary procedures of establishing the item pools for the WT and the IT are detailed.

Item Pools and Samples

Before any testing was done, the test items for both the WT and IT were identified. Since it was desirable that the results of this study be generalizable to the item populations, a pool of 50 representative mathematical problems (Appendix C) for the IT and a pool of 165 items (Appendix D) for the WT were created through the procedures proposed in Chapter IV. The WT was created by randomly selecting 20 items (nos. 47, 89, 46, 28, 159, 29, 48, 25, 2, 91, 30, 23, 36, 12, 33, 75, 101, 96, 61 and 94 in that order) from the appropriate pool.

Six mathematical problems (nos. 10, 28, 34, 36, 4, and 47) were randomly selected for the IT. However, observations of the subjects' text by the investigator and comments by their teachers necessitated changes in the IT sample. Items 28 and 36 both dealt with volumes. The teachers stated that the students had not been acquainted with the topic and an inspection of the subjects' textbook verified that they had not reached the treatment of volumes. Thus items 28 and 36 were randomly replaced by problems 2 and 31 respectively. The WT and the

BEST COPY AVAILABLE

IT were then ready to be administered.

Population

The study was conducted at an elementary, parochial school located in the west central part of Madison. Its 435 students came mainly from middle to upper middle class families of white collar workers and professionals. At the time the study was conducted, the school operated grades 1-8 in a 14 year old two-story building with 16 classrooms.

The mathematics program in grades 5-8 was partially individualized with the students working at their own pace through the Scott-Foresman (Van Engen, et al, 1969) series. The two seventh-grade mathematics teachers supplemented the basic text with handouts, activities, and projects. All of the 63 seventh-graders were included in the first part of the study.

Written Test Administration

On February 24, 1974, the two mathematics teachers administered the WT to 61 subjects. Two students who were absent took the test one week later. Testing procedures had been discussed at preparatory conferences, so each teacher followed the same routine. The two mathematics classes met at consecutive hours in the morning, so it was assumed that the second group did not receive any test information from the earlier group. Each class had approximately 40 minutes to complete the test.

BEST COPY AVAILABLE

The procedures for administering the WT did not directly follow the plans. Originally, half of the subjects would have taken the WT after the IT. The realities of school operations dictated otherwise. First, the teachers both preferred to give the test on the same day. Second, if a random selection of subjects were given the WT at a later date, then each class would be disturbed twice. The investigator did not wish to infringe upon class time more than was necessary.

Another change in plans occurred in the WT item format. Originally, the 20 items were to be presented in random order to avoid a sequence effect. Such an arrangement would have required that each of the 63 tests be typed individually. To permit rapid production of the written tests, the original plan was abandoned and the 20 items were all presented in the same order.

The changes in the WT order and format should not significantly affect the experimental design and statistical results of this study. The possible order effect of the 20 items was disregarded since the items were assumed to be equally representative. Furthermore, though the order effect may have affected the absolute performance of subjects, this study considered the ranking generated by the scores and it was assumed that order effect of the items had negligible effect on the ranks of subjects. The order effect of giving the WT before the IT was also assumed to be minimal because of the emphasis on rank rather than absolute performance of the subjects. The decision

to administer all the written tests at the same time avoided an undesirable complication. If one-half of the students took the WT at a later time, then a time-learning factor could enter the differences in the WT scores of the earlier and later groups.

After the written tests had been completed, the investigator visited the classrooms to discuss the WT with the subjects and to seek their cooperation in arranging the thinking aloud interviews. The subjects had performed poorly (5.6 average number correct) on the WT and had to be reassured that their inability to answer the items correctly may have been due to the items themselves. It was further indicated that the investigator was as interested in the source of their difficulties as he was in the procedures students used to arrive at correct solutions. Each individual was encouraged to accept an invitation to participate in the interviews whether or not he felt that he had done well on the WT. (They were not told their results on the WT.) The procedures for drawing a sample of the seventh graders to participate in the thinking aloud interviews were implemented after the investigator visited the classrooms.

The Interview Sample

In order to heed Kilpatrick's concern for the pressure placed upon subjects in interview situations, it had been decided to choose subjects with at least average mathematical ability for the interviews. But no recent test scores were available with which to classify students.

Following a citywide trend to discourage standardized testing programs, the school had resorted to its own achievement tests which were administered only to students whose status in a subject area was unknown. Thus, not all of the seventh graders had taken the mathematics achievement test. The philosophy of the school did not encourage the administration of an extra test to determine student status, so prior to the WT the mathematics teachers were asked to identify the students who were achieving at least average in their class. Thirty-one average or above average subjects were identified.

It had been planned that a sample of 30 students who were at least average in mathematics ability would be chosen for the interviews of the main study. Since only 31 potential subjects were identified by their teachers no random selection was made: all 31 students were invited to participate in the interviews. The invitations were extended after the WT discussion via a letter mailed to the parents and, within two weeks, 31 affirmative replies were received.

The Interview Arrangements

The video taped interviews were scheduled for the last week in February with the audio taped interviews scheduled for the week after. The staff had graciously agreed to dismiss each subject whenever he was needed, so a schedule covering the entire school day was organized and distributed. Only teachers received schedules: the students were

BEST COPY AVAILABLE

not informed of others who were participating in the program nor the times individuals were scheduled. A student was simply summoned from his class at the time he was needed.

Fifteen of the 31 subjects had been randomly selected to be video taped. A later decision included a sixteenth subject. The video taped interviews were conducted in a mobile unit belonging to the Wisconsin Research and Development Center. The unit, which resembles a mobile home striped of all interior furnishing, was parked on a paved playground adjacent to the school. The observer (the investigator acted as the observer in all the thinking aloud interviews) personally summoned a subject from the classroom, escorted him to the unit, and seated him at a small square table facing the front of the unit. A front view camera stood approximately six feet away from the table and a second camera had been pre-focused over the subject's left shoulder so that the tests and writing were visible. A microphone had been placed near the back edge of the table. The observer sat along the adjacent side on the left of the subject. In a room at the rear of the unit, an operator sat unseen next to a video tape recorder and a monitor. According to prearranged directions, he started recording, made decisions concerning which camera view to record, and stopped recording.

The audio taped interviews were conducted in a different setting. A meeting room in the school basement was used. Subjects were personally summoned by the observer and seated near one end of a long,

BEST COPY AVAILABLE

narrow table so that the observer could sit on their right along the narrow side. A small microphone was placed on the table about two feet in front of the subject. The tape recorder was located on a chair next to the observer so that he could operate the machine while conducting the interviews. The disparity in the physical arrangements of the audio and video taped interviews may account for some of the behavioral differences noticed during the sessions.

The Interview Proceedings

Each video taped interview proceeded as scheduled. Subjects were summoned by the observer and escorted to the mobile unit. The observer began a conversation as he and the subject left the classroom. A restatement of the purpose of the interviews was included in an attempt to set the subject at ease. A conversational tone was continued as the subject entered the mobile unit, was introduced to the recording operator, and was seated at the table. During the reading of the instructions, the observer interrupted after each paragraph and rule to explain or exemplify. The unimportance of getting the correct solutions or of working rapidly was emphasized after the subject read the first paragraph in the instructions and an example of how thinking aloud might sound was given after the second paragraph. Each rule was briefly discussed after it was read and the reasons for not discussing the content of the problems with friends was stressed. After the adjustment period, the subject was

directed to the sample problem. After the subject read and attemped to solve the sample problem, the observer commented upon the subject's thinking aloud procedures, encouraging him to be more vocal if necessary. The subject then began the first of the six problems which were presented in random order. After the subject completed his efforts on a problem, he was asked if he needed a break or was ready to continue. After the sixth problem, a short discussion of the problems and the student's feelings and previous experience usually followed. At the conclusion of the interview, the subject was thanked for his (her) cooperation and escorted back to the classroom.

The audio taped interviews were conducted essentially the same way as the video taped sessions. The necessity for the observer to operate the equipment while conducting the interviews was one departure from the routine. The tape recorder was kept on a chair where it was not conspicuous to the subject and the observer's actions of operating the recorder did not appear to distract the subjects.

A second departure from the routine resulted from the behaviors exhibited by subjects during the video taped interviews. It had become apparent that subjects who were video taped in the mobile unit were nervous. The direct indications were exhibited through physical actions as pencil tapping on the table, shifting feet, and

slightly trembling hands. Subtle indicators were a practice of subjects to combine data without any apparent reason (noted in the pilot study), a general haste in solving problems, and a tendency to opt for abandoning a problem without making a serious attempt to solve it. At the risk of disturbing the design and assumptions for measuring audio and video taping differences, the investigator re-emphasized rule 3 of the instructions by asking students to make an attempt to solve a perplexing problem if they had some knowledge or ideas. They were permitted to stop if they felt that they could make no progress, but it was obvious that the audio taped subjects made more efforts than video taped subjects. Some of the extra effort may have been due to the encouragement of the investigator or it may have been a natural result of another factor; the subjects in the audio taped interviews did not appear as nervous as those who were video taped. It is likely that the familiar setting of an often used school meeting room caused less apprehension than did the novel mobile unit and obvious video taping cameras. The nervousness of the video taped subjects may be attributed to logical causes; the novelty of the setting, the usual apprehension of subjects participating in a study particularly in an individual confrontation, or the particular sample of subjects.

The investigator observed one obstacle that was not mentioned by Lucas or Kilpatrick; the seventh graders had difficulty thinking aloud. One subject would mentally do part of a problem, then write

it down and explain his reasoning (a form of retrospection). When encouraged to verbalize as he thought, the subject remarked that he could not do both simultaneously and that verbalizing interfered with his thinking. The extent of the thinking aloud difficulties is reported and analyzed in Chapter VI.

The fifteen audio taped interviews were concluded on schedule with adjustments made on two occasions when subjects were absent. A week after the tapings were completed, the investigator returned during the mathematics classes to thank the students and teachers for their cooperation, to discuss the IT problems, and to reward everyone with an unexpected treat.

Summary of Chapter V

The changes in routine and additional questions suggested by the pilot study have been identified. The main study was conducted by the revised plans, but produced situations requiring further modifications. Observations of subjects during the interviews noted differences in subjects' behaviors between the two taping procedures. Explanations for some of the differences were offered in Chapter V. Further analysis and the data resulting from the IT and WT are reported in Chapter VI.

☆ U.S.GOVERNMENT PRINTING OFFICE:1974— 651-298/34

BEST COPY AVAILABLE

National Evaluation Committee

Helen Bain
Past President
National Education Association

Lyle E. Bourne, Jr.
Institute for the Study of Intellectual Behavior
University of Colorado

Sue Buel
Dissemination and Installation Services
Northwest Regional Educational Laboratory

Francis S. Chase
Professor Emeritus
University of Chicago

George E. Dickson
College of Education
University of Toledo

Chester W. Harris
Graduate School of Education
University of California

Hugh J. Scott
Consultant
National Evaluation Committee

H. Craig Sipe
Department of Instruction
State University of New York

G. Wesley Sowards
Dean of Education
Florida International University

Joanna Williams
Professor of Psychology and Education
Columbia University

Executive Committee

William R. Bush
Director, Program Planning and Management
Deputy Director, R & D Center

M. Vere DeVault
Professor
School of Education

Herbert J. Klausmeier
Principal Investogator
R & D Center

Joel R. Levin
Principal Investigator
R & D Center

Donald N. McIsaac
Associate Dean, School of Education
University of Wisconsin

Richard A. Rossmiller, Committee Chairman
Director
R & D Center

Len VanEss
Associate Vice Chancellor
University of Wisconsin–Madison

Dan Woolpert
Director, Management Systems
R & D Center

Faculty of Principal Investigators

Vernon L. Allen
Professor
Psychology

B. Dean Bowles
Associate Professor
Educational Administration

Frank H. Farley
Associate Professor
Educational Psychology

Marvin J. Fruth
Associate Professor
Educational Administration

John G. Harvey
Associate Professor
Mathematics

Frank H. Hooper
Associate Professor
Child Development

Herbert J. Klausmeier
V. A. C. Henmon Professor
Educational Psychology

Gisela Labouvie
Assistant Professor
Educational Psychology

Joel R. Levin
Associate Professor
Educational Psychology

L. Joseph Lins
Professor
Institutional Studies

James Lipham
Professor
Educational Administration

Wayne Otto
Professor
Curriculum and Instruction

Robert Petzold
Professor
Curriculum and Instruction

Thomas A. Romberg
Associate Professor
Curriculum and Instruction

Dennis W. Spuck
Assistant Professor
Educational Administration

Richard L. Venezky
Associate Professor
Computer Science

Larry M. Wilder
Assistant Professor
Communication Arts